Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 64(2) 2010 261-272

Smooth Estimation of Survival and Density Functions for Stationary Associated Sequences: Some Recent Developments

Yogendra P. Chaubey[1*] **and Isha Dewan**[2]
[1]*Concordia University, Montreal, Quebec, Canada*
[2]*Indian Statistical Institute, Stat Math Unit, Delhi, India*

Received 02 February 2010; Revised 26 May 2010; Accepted 26 May 2010

SUMMARY

Consider a sequence of stationary non-negative associated random variables with common marginal density $f(x)$. Here we present a review of recent developments for estimating the density f and the corresponding survival function by smoothing the empirical survival function studied in Bagai and Prakasa Rao (1991). These are contrasted with other estimators available for non-negative i.i.d. data.

Keywords: Associated sequence, Asymmetric kernel estimator, Strong consistency, Survival function.

1. INTRODUCTION

Consider a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and a sequence $\{X_n, n \geq 1\}$ of random variables defined on it. A finite collection of random variables $\{X_1, ..., X_n\}$ is said to be associated if for every pair of functions $h(\mathbf{x})$ and $g(\mathbf{x})$ from $\mathbb{R}^n$ to $\mathbb{R}$, which are nondecreasing componentwise,

$$\text{Cov}(h(\mathbf{X}), g(\mathbf{X})) \geq 0$$

whenever it is finite, where $\mathbf{X} = (X_1, X_2, ..., X_n)$. An infinite sequence $\{X_n, n \geq 1\}$ of random variables is said to be associated if every finite subset is associated.

The concept of associated random variables was introduced by Esary *et al*. (1967) in the context of reliability studies. Now these are of considerable interest in many areas of statistical enquiry. They are prominently featured in theory of life testing and reliability (Barlow and Proschan 1981), statistical physics (Newman 1980, 1983) and percolation theory (Cox and Grimmet 1984). The reader may be referred to Newman (1984) for a wealth of results on asymptotic theory involving associated random variables and to Roussas (1999) and Prakasa Rao and Dewan (2001) for an extensive review of several probabilistic and statistical results concerning associated random variables.

Suppose $\{X_n, n \geq 1\}$ is a stationary sequence of associated random variables, where stationarity is meant to indicate that the joint finite dimensional distributions are invariant to translation of the indices by an integer k. Suppose that the density of X_1 exists that is denoted by f. We denote by F and $S = 1 - F$, respectively, the distribution function and the survival function of X_1. In this paper we concentrate on nonparametric estimation of these functions. When the observations are i.i.d., several authors have suggested density estimators based on kernels, histogram methods, orthogonal functions, etc. (see, e.g. Prakasa Rao 1983, Devroye 1989 and Wand and Jones 1995).

Density estimation is not uncommon to the field of agricultural statistics. For example, Qaim (2003) shows a graph of yield density for cotton in India of biotechnology-hybrid and non-biotechnology counterparts (see their Figure 1, pp. 2119). See also Ferreyra *et al*. (2001) where density estimation is used in assessing maize production risk associated with

[*] *Corresponding author* : Y.P. Chaubey
E-mail address : chaubey@alcor.concordia.ca

climate variability. Agricultural field trials are also considered spatially dependent. Hallin *et al.* (2001) model such data by a spatial linear process that is a particular case of positive association considered in this paper.

The most commonly used estimator of the density function is the kernel estimator (Rosenblatt 1956, Parzen 1962) given by

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} k((x - X_i)/h_n) \qquad (1.1)$$

where k is the kernel function, which is generally a symmetric density function with mean zero and variance 1 and h_n is the bandwidth.

Bagai and Prakasa Rao (1995) showed that such kernel type density estimator can be extended to associated sequences and showed that the resulting estimator is strongly consistent, pointwise as well as uniformly, over certain sets. Roussas (1991) studied strong uniform consistency of kernel estimates of r – *th* order derivative of f under some regularity conditions on the kernel and the band- width. In this paper they also considered estimators of the failure rate function $r(x) = f(x) / S(x)$ and the survival function $S(x)$ based on the kernel estimator $f_n(x)$. Later Dewan and Prakasa Rao (1999) considered a general method of nonparametric density estimator in this context given by

$$f_n^{DR}(x) = \frac{1}{n}\sum_{i=1}^{n}\phi_n(x, X_i) \qquad (1.2)$$

where $\phi_n(x, y)$, $n = 1, 2, ...$ is a sequence of Borel-measurable functions defined on $\mathbb{R}^2$ that integrate to 1 (with respect to x). This estimator is a generalization of the histogram type estimator, the kernel type density estimator and the density estimator obtained by the method of orthogonal series.

Note that in the context of life-testing and reliability studies and other areas, the underlying random variables are defined on the nonnegative support. And in such a case, the direct use of kernel density estimator, where k is a usually asymmetric function, is not desirable (as noticed by Silverman 1986). Since this estimator assigns positive mass for x in the interval $(-\infty, 0)$, a properly normalized truncated kernel estimator has to be used to get a proper density estimator. This may not be a satisfactory solution, especially, when the true density takes value zero at $x = 0$. This point is illustrated in Fig. 1 that gives the true density of a gamma distribution with shape parameter 2 and scale parameter 1 along with the kernel estimator, and its truncated version for a sample size 100. Here the kernel used is Gaussian and the bandwidth h is selected using rule of thumb estimator [see Silverman (1986, page 48, Eq. (3.31))]. We see here that the density near $x = 0$ is grossly inaccurate.

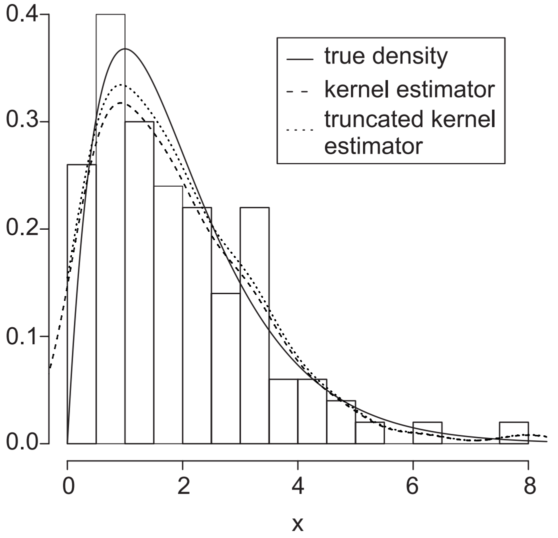

Fig. 1. Kernel density estimator for a sample of gamma density with shape parameter = 2, kernel = Gaussian, sample size = 100.

Several direct methods are now available to deal with this problem for the i.i.d. data. Bagai and Prakasa Rao (1996) suggested the use of a kernel k which is defined only on the positive part of the real line. However, this approach makes use of only the first r order statistics for the value of x in $[X_{r:n}, X_{(r+1):n})$ where $X_{i:n}$ denotes the i^{th} order statistic from the random sample $\{X_1, ... , X_n\}$.

Gawronski and Stadmüler (1980, 1981) have investigated the asymptotic properties of the density estimator based on smoothing histogram using Poisson weights motivated by a result in Feller (1965, Lemma 1, pp. 229). Chaubey and Sen (1996), independently suggested a truncated version of this density estimator that was found to be unsuited for estimation of the mean residual life function (see Chaubey and Sen 1999). A subsequent modification resulted in the same estimator as proposed by Gawronski and Stadmüler (1980). This has been extended recently, to the case of associated data in Chaubey and Dewan (2009a). This method uses the whole data in contrast to the method of Bagai and Prakasa Rao (1996), however, it may not be quite appropriate at the lower most boundary for removing the bias. This problem was addressed in Chaubey *et al.* (2007) for i.i.d. sequences. They used a generalization of Hille's lemma along with a perturbation idea. This method is appropriate for dealing with the boundary bias problem at zero and fits in with the generalized estimator $f_n^{DR}(x)$. This has been studied recently by Dewan and Chaubey (2009b) for the case of non-negative associated sequences.

The class of estimators introduced by Chen (2000) using Gamma kernels and those by Scaillet (2004) using inverse Gaussian and reciprocal inverse Gaussian kernels, proposed for smooth density estimation for non-negative i.i.d. data may be similarly adapted for associated sequence also. These are not highly desirable in comparison to the estimator given in Chaubey *et al.* (2007), as their variances seem to blow up at $x = 0$, and their behavior for estimating densities that may not be zero at $x = 0$ is not clear. Chen's estimator has been more thoroughly investigated recently by Zhang (2010). The author reports that "our study finds that the Gamma kernel estimator at $x = 0$ is actually the reflection estimator when the double exponential kernel is used and is only boundary problem free when the estimated density has a shoulder at $x = 0$ (i.e., the first derivative of the density at $x = 0$ is zero). For densities not satisfying the shoulder condition, we show that the gamma kernel estimator has a severe boundary problem and its performance is inferior to that of the boundary kernel estimator."

A qualitative comparison of new estimators for the sample used in Fig. 1 is illustrated in Fig. 2. The value of λ_n used here is obtained from the working rule $\lambda_n = n/\max(X_1, ..., X_n)$ [see Chaubey and Sen (2009)]. Both the density estimators, the one based on Poisson weights (as in (2.12) with $\lambda_n = 12.52$) and the Gamma kernel estimator (as in (2.14), $v_n = h$ and $\varepsilon_n = 0$), are seen to avoid the boundary value problem near zero. These smoothing constants may be obtained using cross-validation methods as discussed in Chaubey and Sen (2009) for the Poisson weight estimator and in Chaubey *et al.* (2007) for the asymmetric kernel estimator, however, we have used their rough values just to demonstrate the qualitative aspect of the new estimators.

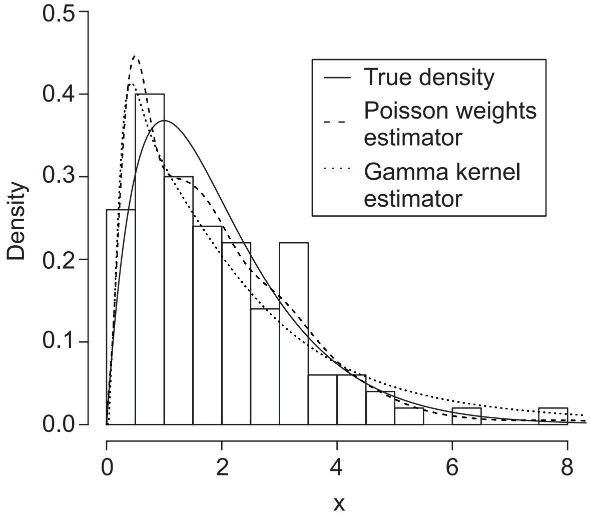

Fig. 2. Poisson weight density estimator and Gamma kernel density estimator for a sample of gamma density with shape parameter = 2, sample size = 100.

The purpose of the present article is to review some of these recent developments for smooth estimation of density and survival functions for non-negative associated sequences. In Section 2, we present the generalized smoothing lemma to motivate the new estimators of the density and survival functions as studies in Chaubey and Dewan (2009a, b). Section 3 presents some properties of the resulting smooth survival function estimator and Section 4 presents those for the smooth density estimator. The final section contains some concluding remarks.

2. THE GENERALIZED SMOOTHING LEMMA AND SMOOTHING OF THE EMPIRICAL DISTRIBUTION FUNCTION

Before we present the lemma that is the key to discussions of the results in this paper, we start with results of Bagai and Prakasa Rao (1991) concerning the empirical survival function $S_n(x)$ based on a sequence of associated random variables $\{X_n, n \geq 1\}$ given by

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_j(x) \qquad (2.1)$$

where

$$Y_j(x) = \begin{cases} 1 & \text{if } X_j > x \\ 0 & \text{otherwise} \end{cases}$$

Note that when the underlying random variable is continuous, the empirical survival function needs to be modified, in such a way as to produce smooth estimator. Suppose that $F(x)$, is absolutely continuous, so that $F(x)$ admits a density function $f(x)$. With this objective any non-parametric density estimator for $f(x)$ may be properly integrated to produce the desired estimator. For example suppose we consider the kernel density estimator (considered by Dewan and Prakasa Rao 1999),

$$f_n(x) = \frac{1}{nh_n} \sum_{j=1}^{n} k\left(\frac{x - X_j}{h_n}\right) \tag{2.2}$$

where $k(.)$ is a suitable kernel and $\{h_n, n \geq 1\}$ is a bandwidth sequence,then a nonparametric (smooth) estimator of the survival function $S(x)$ is given by

$$S_{nK}(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K_n(x, X_j) \tag{2.3}$$

where

$$K_n(x, y) = \int_x^{\infty} k\left(\frac{s - y}{h_n}\right) \mathrm{d}s \tag{2.4}$$

As indicated in the introduction, kernel density estimator (and hence the resulting survival function estimator) may not be universally appropriate, suitable modifications may be necessary in specific situations. Chaubey and Dewan (2009a,b) have investigated the adaptation of smoothing the empirical survival function as presented in Chaubey and Sen (1996) and Chaubey *et al.* (2007) to associated sequences. These are motivated by the following lemma, which is a slight variation of Lemma 1 given in Feller (1965, §VII.1).

Lemma 2.1. Let u be any bounded and continuous function. Let $G_{x,n}$, $n = 1, 2, \ldots$ be a family of distributions with mean $\mu_n(x)$ and variance $h_n^2(x)$ then we have as $\mu_n(x) \to x$ and $h_n(x) \to 0$

$$\tilde{u}(x) = \int_{-\infty}^{\infty} u(t) \mathrm{d}G_{x,n}(t) \to u(x) \tag{2.5}$$

The convergence is uniform in every subinterval in which $h_n(x) \to 0$ and u is uniformly continuous.

This generalization may be adapted for smooth estimation of the distribution function by replacing $u(x)$ by the empirical distribution function $F_n(x) = 1 - S_n(x)$ as given below

$$\tilde{F}_n(x) = \int_{-\infty}^{\infty} F_n(t) \mathrm{d}G_{x,n}(t) \tag{2.6}$$

Strong convergence of $\tilde{F}_n(x)$ parallels to that of the strong convergence of the empirical distribution function as stated in the following theorem.

Theorem 2.1. If $h \equiv h_n(x) \to 0$ for every fixed x as $n \to \infty$ we have

$$\sup_x | \tilde{F}_n(x) - F(x) | \overset{a.s.}{\to} 0 \tag{2.7}$$

as $n \to \infty$.

Technically, $G_{x,n}$ can have any support but it may be prudent to choose it so that it has the same support as the random variable under consideration; because this will get rid of the problem of the estimator assigning positive mass to undesired region.

For $\tilde{F}_n(x)$ to be a proper distribution function, $G_{x,n}(t)$ must be a decreasing function of x, which can be shown using an alternative form of $\tilde{F}_n(x)$:

$$\tilde{F}_n(x) = 1 - \frac{1}{n} \sum_{i=1}^{n} G_{x,n}(X_i) \tag{2.8}$$

In addition to being computationally attractive, this form provides an insight into the usual kernel estimator. This also leads us to propose a smooth estimator of the density given by

$$\tilde{f}_n(x) = \frac{\mathrm{d}\tilde{F}_n(x)}{\mathrm{d}x}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}x} G_{x,n}(X_i) \tag{2.9}$$

Remark 2.1. It is interesting to note that the form of the above density estimator is similar to the general method proposed by Dewan and Prakasa Rao (1999), with

$$\phi_n(x, y) = -\frac{\mathrm{d}}{\mathrm{d}x} G_{x,n}(y)$$

Densities with Non-Negative Support

Smoothing based on Poisson weights

Considering $G_{x,n}$ defined by a Poisson distribution attaching weights $p_k(\lambda_n x)$, $k = 0, 1, 2, \ldots$ to the lattice points k/λ_n, in (2.5), we obtain Hille's approximation lemma. Replacing $u(x)$ by $S_n(x)$, we get a smooth estimator of survival function as considered in Chaubey and Dewan (2009a):

$$\tilde{S}_n^P(x) = \sum_{k=0}^{\infty} S_n\left(\frac{k}{\lambda_n}\right) p_k(\lambda_n x) \qquad (2.10)$$

$$p_k(\mu) = e^{-\mu}\frac{\mu^k}{k!},\ k = 0, 1, \ldots, \qquad (2.11)$$

and λ_n is a sequence (possibly stochastic) of constants tending to infinity as $n \to \infty$. The corresponding estimator of the density, that was originally proposed in Gawronski (1980) in the context of *i.i.d.* data, is given by

$$\begin{aligned}\tilde{f}_n^P(x) &= -\frac{\mathrm{d}}{\mathrm{d}x}\tilde{S}_n(x) \\ &= \lambda_n \sum_{k=0}^{\infty} p_k(\lambda_n x) w(k, \lambda_n) \qquad (2.12)\end{aligned}$$

where

$$w(k, \lambda_n) = \left[S_n\left(\frac{k}{\lambda_n}\right) - S_n\left(\frac{k+1}{\lambda_n}\right)\right]$$

Smoothing based on Asymmetric Kernels

Using the representation (2.8), Chaubey and Dewan (2009b) proposed the following estimators of the distribution and density functions with support $[0, \infty)$, generalizing the estimator based on Poisson weights. Let $Q_v(x)$ represent a distribution on $[0, \infty)$ with mean 1 and variance v^2, then an estimator of $F(x)$ is given by

$$F_n^+(x) = 1 - \frac{1}{n}\sum_{i=1}^{n} Q_{v_n}\left(\frac{X_i}{x}\right) \qquad (2.13)$$

where $v_n \to 0$ as $n \to \infty$. Obviously, this choice uses $G_{(x,n)}(t) = Q_{v_n}(t/x)$ which is a decreasing function of x.

This leads to the following density estimator

$$\frac{\mathrm{d}}{\mathrm{d}x}(F_n^+(x)) = \frac{1}{nx^2}\sum_{i=1}^{n} X_i q_{v_n}\left(\frac{X_i}{x}\right)$$

where $q_v(.)$ denotes the density corresponding to the distribution function $Q_v(.)$.

However, the above estimator may not be defined at $x = 0$, except in cases where $\lim_{x\to 0} \frac{\mathrm{d}}{\mathrm{d}x}(F_n^+(x))$ exists. Moreover, this limit is typically zero, which is acceptable only when we are estimating a density f with $f(0) = 0$.

Hence in view of the more general case where $0 \le f(0) < \infty$, Chaubey and Dewan (2009b) adapt the following *perturbed* version of the above density estimator:

$$f_n^+(x) = \frac{1}{n(x+\varepsilon_n)^2}\sum_{i=1}^{n} X_i q_{v_n}\left(\frac{X_i}{x+\varepsilon_n}\right),\ x \ge 0 \qquad (2.14)$$

where $\varepsilon_n \downarrow 0$ at an appropriate (sufficiently slow) rate as $n \to \infty$. In the sequel, we illustrate our method by taking $Q_v(.)$ to be the Gamma ($\alpha = 1/v^2$, $\beta = v^2$) distribution function.

Next we present a comparison of our approach with some existing estimators.

Kernel Estimator. The usual kernel estimator is a special case of the representation given by Eq. (2.9), by taking $G_{x,n}(.)$ as

$$G_{x,n}(t) = K\left(\frac{t-x}{h}\right) \qquad (2.15)$$

where $K(.)$ is a distribution function with mean zero and variance 1.

Transformation Estimator of Wand *et al.* The well known logarithmic transformation approach of Wand *et al.* (1991) leads to the following density estimator:

$$\tilde{f}_n^{(L)}(x) = \frac{1}{nh_n x}\sum_{i=1}^{n} k\left(\frac{1}{h_n}\log(X_i/x)\right)$$

where $k(.)$ is a density function (kernel) with mean zero and variance 1. This is easily seen to be a special case of Eq. (2.9), taking $G_{x,n}$ again as in Eq. (2.15) but applied to log x. This approach, however, creates problem at the boundary which led Ruppert and Marron (1994) to propose modifications that are computationally intensive.

Estimators of Chen and Scaillet. Chen's (2000) estimator is of the form

$$\hat{f}_C(x) = \frac{1}{n}\sum_{i=1}^{n} g_{x,n}(X_i)$$

where $g_{x,n}(.)$ is the Gamma($\alpha = a(x, b)$, $\beta = b$) density with $b \to 0$ and $ba(x, b) \to x$. This also can be motivated from Eq. (2.5) as follows: Take $u(t) = f(t)$ and note that the integral $\int f(t)g_{x,n}(t)\mathrm{d}t$ can be estimated by $n^{-1} \sum_{i=1}^{n} g_{x,n}(X_i)$. This approach controls the boundary bias at $x = 0$; however, the computation of integrated mean squared error (IMSE) is not tractable. Moreover, estimators of derivatives of the density are not easily obtainable because of the appearance of x as argument of the Gamma function. Scaillet's (2004) estimators replace the Gamma kernel by inverse Gaussian (IG) and reciprocal inverse Gaussian (RIG) kernels. These estimators are more tractable than Chen's; however, they assume value zero at $x = 0$ which may not be desirable in all cases.

3. ASYMPTOTIC PROPERTIES OF SMOOTH ESTIMATORS OF THE SURVIVAL FUNCTION

It is useful to describe the asymptotic properties of $S_n(x)$ before considering those for the smooth estimators. The following theorems regarding weak and strong convergence of $S_n(x)$ were established in Bagai and Prakasa Rao (1991).

Theorem 3.1. [Bagai and Prakasa Rao (1991)] Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of associated random variables with bounded continuous density for X_1. Assume that, for some $r > 1$

$$\sum_{j=n+1}^{\infty} \{\mathrm{Cov}\ (X_1, X_j)\}^{1/3} = O(n^{-(r-1)}) \tag{3.1}$$

Then

$$S_n(x) \to S(x) \text{ a.s. as } n \to \infty$$

Theorem 3.2. [Bagai and Prakasa Rao (1991)] Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of random variables satisfying the conditions of Theorem 3.1. Then for any compact subset $J \subset R$

$$\sup[|S_n(x) - S(x)\,| : x \in J] \to 0 \text{ a.s. as } n \to \infty$$

Yu (1993) established uniform strong consistency assuming $S(x)$ to be continuous.

Theorem 3.3. [Yu (1993)] Let $\{X_n, n \geq 1\}$ be a sequence of associated random variables having the same continuous marginal distribution function $F(x)$ for X_n, $n \geq 1$. If

$$\sum_{n=1}^{\infty} \frac{1}{n^2}\ \mathrm{Cov}(X_n, T_{n-1}) < \infty \tag{3.2}$$

where $T_n = \sum_{j=1}^{n} X_j$, then, as $n \to \infty$

$$\sup_{-\infty<x<\infty} |\, S_n(x) - S(x)\,| \to 0 \text{ a.s.}$$

If the sequence $\{X_n, n \geq 1\}$ is stationary, then the condition in Eq. 3.2 can be relaxed to

$$\sum_{i=1}^{n} \frac{1}{n}\ \mathrm{Cov}(X_n, X_i) < \infty \tag{3.3}$$

Bagai and Prakasa Rao (1991) also obtained the asymptotic law of $S_n(x)$ that is quoted below.

Theorem 3.4. [Bagai and Prakasa Rao (1991), Theorem 3.3] Let $\{X_n, n \geq 1\}$ be a stationary sequence of associated random variables with bounded continuous density for X_1 and survival function $S(x)$. Suppose that $n \to \infty$ and

$$\sum_{j=2}^{\infty} \{\mathrm{Cov}(X_1, X_j)\}^{1/3} < \infty \tag{3.4}$$

Then, for all x such that $0 < S(x) < 1$, $\sqrt{n}\,[S_n(x) - S(x)]/\sigma(x)$ converges in distribution to a standard normal distribution, where

$$\sigma^2(x) = S(x)[1 - S(x)] + 2\sum_{j=2}^{\infty} \{P[X_1 > x, X_j > x] - S^2(x)\}$$

Chaubey and Sen (2009a,b) established the following version of Glivenko-Cantelli theorem for

$\tilde{S}_n(x)$ similar to the one obtained in Bagai and Prakasa Rao (1991) for $S_n(x)$. Note that here and elsewhere $\tilde{S}_n(x)$ will be used to denote either of $\tilde{S}_n^P(x)$ or $\tilde{S}_n^+(x)$.

Theorem 3.5. Let $\{X_n, n \geq 1\}$ be a stationary sequence of associated random variables with bounded continuous density for X_1. Assume that, for some, $r > 1$,

$$\sum_{j=n+1}^{\infty} \{\mathrm{Cov}(X_1, X_j)\}^{1/3} = O(n^{-(r-1)})$$

and $\max(X_1, X_2, \ldots, X_n) \to \infty$. Then for $\lambda_n \to \infty$ $\|\tilde{S}_n - S\| = \sup\{|\tilde{S}_n(x) - S(x)| : x \in \mathbf{R}^+\} \to 0$ a.s., as $n \to \infty$. (3.5)

Chaubey and Dewan (2009a) use Bahadur-type representation established by Ekisheva (2001) to establish the order of closeness of $\tilde{S}_n^P$ and S_n, however we require a stronger condition than used in Bagai and Prakasa Rao (1991).

Theorem 3.6. Suppose that $\lambda_n = O(n)$ as in Theorem 2.1, whenever $f(x)$ is absolutely continuous with a bounded derivative $f'(.)$ a.e. on $\mathbf{R}^+$, and

$$\sum_{n=1}^{\infty} n^7 \mathrm{cov}(X_1, X_n) < \infty \tag{3.6}$$

$$\|\tilde{S}_n^P - S_n\| = O(n^{-5/8}(\log n)) \text{ a.s. as } n \to \infty \tag{3.7}$$

Remark 3.2. The above theorem can be shown to hold for $S_n^+(x)$, under appropriate conditions on the convergence of v_n and ε_n following similar steps as in Chaubey *et al.* (2007).

The asymptotic law of the smooth estimators is found to be the same as that of the raw estimator as given in the following theorem.

Theorem 3.7. Let $\{X_n, n \geq 1\}$ be a stationary sequence of associated random variables and assume that the conditions in Theorem 3.1 are satisfied. Then, for all x such that $0 < S(x) < 1$, $\sqrt{n}\,[\tilde{S}_n(x) - \mathrm{S}(x)]/\sigma(x)$ converges in distribution to a standard normal distribution, where

$$\sigma^2(x) = S(x)[1 - S(x)]$$
$$+2\sum_{j=2}^{\infty} \{P[X_1 > x, X_j > x] - S^2(x)\}$$

4. ASYMPTOTIC PROPERTIES OF SMOOTH ESTIMATORS OF *f*

Kernel Estimators

Bagai and Prakasa Rao (1995) assumed the following conditions

(A1) $k(.)$ is a bounded density function of bounded variation on $\mathbb{R}$ satisfying (i) $\lim_{|u| \to inf} k(u) = 0$, (*ii*) $\int_{-\infty}^{\infty} u^2 k(u)\mathrm{d}u < \infty$.

(A2) $k(x)$ is differentiable and $\sup_x |k'(x)| \leq c < \infty$.

(B) For all l and $r \geq 0$, $\sum_{j:|l-j|\geq r} \mathrm{Cov}(X_j, X_l) \leq u(r)$, where $u(r) = e^{-\alpha r}$ for some $\alpha > 0$.

and obtained the following expressions for the mean and variance of the kernel density estimator under these conditions:

$$E[f_n(x)] = f(x) - h_n f'(x)\gamma_1 + \frac{h_n^2}{2} f''(x)\gamma_2 + O(h_n^3) \tag{4.1}$$

where

$$\gamma_j = \int_{-\infty}^{\infty} x^j k(x)\mathrm{d}x,\ j = 1, 2$$

$$\mathrm{Var}[f_n(x)] = \frac{1}{nh_n}[f(x)\beta_0 + O(h_n)] + O(1/nh_n^4) \tag{4.2}$$

where

$$\beta_0 = \int_{-\infty}^{\infty} k^2(x)\mathrm{d}x$$

The following theorem concerns the strong point wise convergence of $f_n(x)$.

Theorem 4.1. Let $\{X_n, n \geq 1\}$ be a stationary associated sequence of random variables. Suppose that conditions

(A) and (B) hold. Then for $x \in I$ where I is a compact subset of $\mathbb{R}$,

$$f_n(x) - E|f_n(x)] \to 0 \text{ a.s. as } n \to \inf.$$

The above theorem also gives point wise convergence at continuity points of f.

Corollary 4.1. Under the assumptions of Theorem 4.1, $f_n(x) \to f(x)$ a.s. at continuity points x of f as $n \to \infty$.

Under further assumption on h_n, uniform convergence over a compact subset is obtained.

Theorem 4.2. Let $\{X_n, n \ge 1\}$ be a stationary associated sequence of random variables. Suppose that conditions (A) and (B) hold. Further let

$$h_n^{-4} = O(n^\gamma) \text{ for } \gamma > 0$$

and for some constant $C > 0$

$$|f(x_1) - f(x_2)| \le C\,|x_1 - x_2|,\ x_1, x_2 \in \mathrm{I}$$

then for $x \in I$ where I is a compact subset of $\mathbb{R}$

$$\sup_{x \in I} |f_n(x) - f(x)| \to 0 \text{ a.s. as } n \to \infty$$

General Method of Density Estimator

Dewan and Prakasa Rao's (1999) general density estimator $f_n^{DR}(x)$ was proposed for the case of associated sequences based on similar method given by Foldes and Revesz (1974) for *i.i.d.* case. They studied conditions leading to the exponential rate of convergence for the uniform consistency in probability of the estimator and obtained the following theorem.

Theorem 4.3. Let $\{X_n, n \ge 1\}$ be a stationary associated sequence of random variables with common marginal density function f. Consider the following conditions to hold:

(A1) $|f(x_1) - f(x_2)| \le K|x_1 - x_2|$ if $x_1, x_2 \in [a, b]$ (4.3)

(A2) $\int_{-\infty}^{\infty} |x|^\gamma f(x)\mathrm{d}x < \infty$ for some $\gamma > 0$ (4.4)

Let $\{\phi_n(x, y)\}$ be a sequence of Borel-measurable functions of bounded variation in y for every fixed x. Then

$$\phi_n(x, y) = \phi_{1n}(x, y) - \phi_{2n}(x, y) \tag{4.5}$$

where $\phi_{in}(x)$, $i = 1, 2$ is monotone in y for fixed x.

(B1) Suppose that there exists two positive numbers α and τ and an interval $[c, d]$ containing $[a, b]$ such that for each n the interval $[c, d]$ can be divided into disjoint left closed intervals $I_s^{(n)}$, $s = 1, 2, \ldots$ for which

$$|I_s^{(n)}| \ge \frac{1}{n^\alpha},\ \bigcup_{s=1}^{n} I_s^{(n)} = [c, d] \tag{4.6}$$

and

$$|\phi_n(x_1, y) - \phi_n(x_2, y)| \le n^\tau |x_1 - x_2| \tag{4.7}$$

provided that x_1 and x_2 belong to the same interval $I_s^{(n)}$.

(B2) Suppose that

$$\mathrm{Var}(\phi_{in}(x, X_1)) \le h_n,\ i = 1, 2 \tag{4.8}$$

$$h_n \le \frac{n}{w(n)\log n} \tag{4.9}$$

where $w(n) = \mathrm{O}(n^\beta)$, for some $\beta > 0$, and $w(n) \to \infty$ as $n \to \infty$.

(B3) There exists a positive constant C such that

$$|\phi_{in}(x, y)| \le Ch_n,\ i = 1, 2 \tag{4.10}$$

(B4) Suppose that there exists a $v > 0$ and a sequence of positive numbers $\varepsilon_n \to 0$ such that

$$|\phi_n(x_n, y_n)| \le \varepsilon_n \tag{4.11}$$

whenever

$$|x_n - y_n| > n^\gamma \text{ and } n > n_0(\varepsilon) \tag{4.12}$$

(B5) Suppose that $\phi_{in}(x, y)$, $i = 1, 2$ is differentiable with respect to y such that

$$|\phi'_{in}(x, y)| \le b_n \tag{4.13}$$

where $\phi'_{in}(x, y)$ denotes the derivative with respect to y and there exists $\eta > 0$ such that

$$\frac{b_n}{h_n} = \mathrm{O}(n^\eta) \tag{4.14}$$

(C1) Assume that

$$\int_a^b \phi_n(x, y) f(y)\mathrm{d}y \to f(x) \text{ as } n \to \infty \tag{4.15}$$

uniformly in $[a + \delta, b - \delta]$ for some $\delta > 0$.

(C2) Finally assume that

$$\frac{1}{n}\sum_{j=1}^{n} \mathrm{Cov}(X_1, X_j) = O(e^{-n\theta}) \qquad (4.16)$$

for some $\theta > 3/2$.

Then

$$\Pr(\sup_{a+\delta \le x \le b-\delta} | f_n(x) - f(x)| \ge \varepsilon) \le e^{-\frac{k_1 n}{h_n}} \qquad (4.17)$$

as $n \to \infty$, where k_1 is a positive constant depending on ε, δ and f.

Density Estimator Based on Poisson Weights:

For $\tilde{f}_n^P(.)$, defined by (2.12), Chaubey and Dewan (2009a) established the following theorem on strong consistency.

Theorem 4.4. Let $\{X_n, n \ge 1\}$ be a stationary sequence of associated random variables with bounded continuous density f and derivative f'. Assume that, for some $r > 1$

$$\sum_{j=n+1}^{\infty} \{\mathrm{Cov}(X_1, X_j)\}^{1/3} = O(n^{-(r-1)})$$

and the sequence of the constants $\{\lambda_n, n \ge 1\}$ is such that

$$\sum_{n=1}^{\infty} \left(\frac{\lambda_n^2}{n}\right)^r < \infty$$

we have

$$\sup_{x \in R} | \tilde{f}_n^P(x) \to f(x)| \to 0 \text{ a.s. as } n \to \infty \qquad (4.18)$$

Density Estimator Based on Asymmetric Kernels

The bias and variance of $f_n^+(x)$ is given in Chaubey and Dewan (2009b). The bias is given by:

$$\mathrm{Bias}\left[f_n^+(x)\right] = (x v_n^2 + \varepsilon_n) f'(x) + O\left(v_n^2 + \varepsilon_n\right) \qquad (4.19)$$

For computing the variance Chaubey and Dewan (2009a) made the following assumptions as given in Chaubey *et al.* (2007):

(A1) $\int_0^\infty (q_{v_n}(t))^m \, dt = O\left(v_n^{(-m-1)}\right)$ as $v_n \to 0$ for $1 \le m \le 3$

(A2) $I_2(q) = \lim_{v_n \to 0} v_n \int_0^\infty (q_{v_n}(t))^2 \, dt$ exists

(A3) with $q^*_{m,v_n}(t) = \dfrac{(q_{v_n}(t))^m}{\int_0^\infty (q_{v_n}(u))^m du}$, $1 \le m \le 3$, and as $v_n \to 0$

(i) $\mu_{m,v_n} = \int_0^\infty t q^*_{m,v_n}(t) dt = 1 + O(v_n)$

(ii) $\sigma^2_{m,v_n} = \int_0^\infty (t - \mu_{m,v_n})^2 \, q^*_{m,v_n}(t) dt = O\left(v_n^2\right)$

(iii) $\sup_{0 < v_n < \varepsilon_n} \int_0^\infty t^{4+\delta} q^*_{m,v_n}(t) dt < \infty$ for some $\delta > 0$, $\varepsilon > 0$

(A4) Let $\psi_n(x, y) = y q_{v_n}\left(\frac{y}{x+\varepsilon_n}\right)$

$$\frac{\partial}{\partial y}\psi_n(x, y) = q v_n\left(\frac{y}{x+\varepsilon_n}\right) + \frac{y}{x+\varepsilon_n} q'_{v_n}\left(\frac{y}{x+\varepsilon_n}\right)$$

Suppose $\sup_{x,\, y}\left|\frac{\partial}{\partial y}\psi_n(x, y)\right| \le C$, where C is a positive constant.

(A5) For all l and $r \ge 0$, $\sum_{j:|l-j| \ge r} \mathrm{Cov}(X_j, X_l) \le u(l)$ where $u(r) = e^{-\alpha r}$ for some $\alpha > 0$.

We then have

$$\text{Var } f_n^+(x) = \frac{I_2(q)f(x)}{nv_n(x+\varepsilon_n)} + O\left(\frac{1}{n(x+\varepsilon_n)^4}\right) + o((nv_n)^{-1}) \qquad (4.20)$$

as $v_n \to 0$, $\varepsilon_n \to 0$, $nv_n \to \infty$.

Using above results, the Mean Integrated Squared Error of $f_n^+(x)$ is given by

$$\text{MSE}(\ f_n^+(x)) = \text{Var}\ (\ f_n^+(x)) + \text{Bias}^2\left[f_n^+(x)\right]$$
$$= \frac{I_2(q)f(x)}{nv_n(x+\varepsilon_n)} + \left[(xv_n^2+\varepsilon_n)f'(x)\right]^2 + O\left(\frac{1}{n(x+\varepsilon_n)^4}\right) + o\left(v_n^2+\varepsilon_n\right) + o\left(\left(nv_n^{-1}\right)\right)$$

Thus $f_n^+(x)$ is asymptotically unbiased and weakly consistent for $f(x)$.

It also therefore follows that the mean integrated squared error is

$$\text{MISE}(\ f_n^+(x)) = \int_0^\infty \text{MSE}\left(f_n^+(x)\right)dx$$
$$\simeq \frac{I_2(q)}{nv_n}\int_0^\infty \frac{f(x)}{(x+\varepsilon_n)}dx + \int_0^\infty\left[\left(xv_n^2+\varepsilon_n\right)f'(x)\right]^2 dx + \int_0^\infty \frac{1}{n(x+\varepsilon_n)^4}dx + o\left(v_n^2+\varepsilon_n\right) + o\left(\left(nv_n\right)^{-1}\right)$$

The leading term of integrated mean square error is the asymptotic MISE,

$$\text{AMISE}\left[f_n^+\right] = \frac{I_2(q)}{nv_n}\int_0^\infty \frac{f(x)}{(x+\varepsilon_n)}dx + \int_0^\infty\left[\left(xv_n^2+\varepsilon_n\right)f'(x)\right]^2 dx + \int_0^\infty \frac{1}{n(x+\varepsilon_n)^4}dx$$

Chaubey and Dewan (2009b) proved the following theorem on uniform consistency for the asymmetric kernel estimator (see the theorem below) however, the question of asymptotic normality of the density estimator in the associated case is an open problem.

Theorem 4.5. Let $\{X_n, n \geq 1\}$ be a stationary sequence of associated random variables with bounded continuous density for X_1. Assume that, for some $r > 1$

(B1) $\sum_{j=n+1}^{\infty} \{Cov(X_1, X_j)\}^{1/3} = O\ (n^{-(r-1)})$

(B2) $v_n \to 0$, $\varepsilon_n \to 0$ *as* $n \to \infty$

(B3) $\sup_{x\geq 0}\int_0^\infty \left|\frac{\partial}{dx}\left[g_{x+\varepsilon_{n,n}}(t)\right]\right. = O\left(\left(\frac{\log\log n}{n^{1/2}}\right)\right)$

(B4) $\sup_{u>0, v_n>0} uq_{v_n}(u) < \infty$, and

(B5) $f(.)$ is Lipschitz continuous on $[0,\infty)$.

Then, for any compact set $J \subset R$,

$\sup\ [|\ f_n(x) - f(x)|\ x \in J\ | \to 0$ a.s. as $n \to \infty$.

5. CONCLUSIONS

Here we have reviewed some recent results on non-parametric estimation of the survival function and the density function of a sequence of stationary associated random variables. These may be of special importance in the field of agriculture, where we encounter nonnegative data such as crop yields. Moreover, agricultural data may be considered to be spatially dependent that can be modeled using positive association. It is shown here that the usual Parzen-Rosenblatt estimator, proposed under the i.i.d. setup may still be applicable under positive association in general. Further, it can be appropriately modified using asymmetric kernels for the case of non-negative data. The usefulness of the new estimator is illustrated through an example.

A numerical study to compare various estimators available in literature is being done and will be reported elsewhere. One could also study recursive kernel type estimators for the survival function and the density function in this context. Associatedness is one of many types of dependence structures of interest in statistical literature (see Shaked and Shanthikumar (2007)), *viz,* ϕ-mixing or strong mixing. Estimators considered here may also be studied for these other kinds of dependent sequences and there is already a vast literature on some of these (e.g. kernel type estimators), but we believe that a review of these is beyond the scope of this paper.

ACKNOWLEDGEMENT

The research of the first author was partially supported from the author's Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. The authors are grateful to a referee for some constructive comments that have helped to improve the presentation of this article.

REFERENCES

Bagai, I. and Prakasa Rao, B.L.S. (1991). Estimation of the survival function for stationary associated processes. *Statist. Probab. Lett.,* **12**, 385-391.

Bagai, I. and Prakasa Rao, B.L.S. (1995). Kernel type density and failure rate estimation for associated sequences. *Ann. Inst. Statist. Math.,* **47**, 253-266.

Bagai, I. and Prakasa Rao, B.L.S. (1996). Kernel type density estimates for positive valued random variables. *Sankhyā,* **A57**, 56-67.

Barlow, R.E. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*: *Probability Models.* Holt, Reinhart and Winston, New York.

Chaubey, Y.P. and Dewan, I. (2009a). Smooth estimation of survival and density functions for a stationary associated process using Poisson weights. *Technical Report No. 3/09,* Department of Mathematics and Statistics, Concordia University, Montreal, Canada.

Chaubey, Y.P. and Dewan, I. (2009b). An asymmetric kernel estimator of density function for stationary associated sequences. *Preprint isid/ms/2009/10*, Stat Math Unit, Indian Statistical Institute, New Delhi, India.

Chaubey, Y.P., Sen, A. and Sen, P.K. (2007). A new smooth density estimator for non-negative random variables. *Technical Report* No. 1/07, Department of Mathematics and Statistics, Concordia University, Montreal, Canada.

Chaubey, Y.P. and Sen, P.K. (1996). On smooth estimation of survival and density functions. *Statist. Decisions,* **14**, 1-22.

Chaubey, Y.P. and Sen, P.K. (1999) . On smooth estimation of mean residual life. *Jour. Statist. Plann. Inf.,* **75**, 223-236.

Chaubey, Y.P. and Sen, P.K. (2009). On the selection of the smoothing parameter in Poisson smoothing of histogram estimator: Computational aspects. *Pak. J. Statist.,* **25(4)**, 385-401.

Chen, X.X. (2000). Probability density function estimation using Gamma kernels. *Ann. Inst. Statist. Math.,* **52**, 471-480.

Cox, D.R. and Grimmet, G. (1984). Central limit theorems for associated random variables and the percolation model. *Ann. Probab.,* **12**, 514-528.

Devroye, L. (1989). *A Course in Density Estimation.* Birkhauser, Boston.

Dewan, I. and Prakasa Rao, B.L.S. (1999). A general method of density estimation for associated random variables. *Jour. Nonpar. Statist.,* **10**, 405-420.

Esary, J., Proschan, F. and Walkup, D. (1967). Association of random variables with applications. *Ann. Math. Statist.,* **38**, 1466-1474.

Ekisheva, S.V. (2001). Limit theorems for sample quantiles of associated random sequences. *Fundamental and Applied Mathematics* (In Russian, *Fundamentalnaya I Prikladnaya Matematika*), **7**, 721-734.

Feller, W. (1965). *An Introduction to Probability Theory and its Applications, Vol. II.* John Wiley & Sons, New York.

Foldes, A. and Revesz, P. (1974). A general method of density estimation. *Studia Sc. Math. Hungar.,* **9**, 82-92.

Ferreyra, R.A., Podesta, G.P., Messina, C.D., Letson, D., Dardanelli, J., Guevara, E. and Meira, S. (2001). A linked-modeling framework to estimated maize production risk associated with ENSO-related climate variability in Argentina. *Agric. Forest Meteorology,* **107**, 177-192.

Gawronski, W. and Stadmüler, U. (1980). On density estimation by means of Poisson's distribution. *Scand. J. Statist.,* **7**, 90-94.

Gawronski, W. and Stadmüler, U. (1981). Smoothing of histograms by means of lattice- and continuous distributions. *Metrika,* **28**, 155-164.

Hallin, M., Lu, Z. and Lanh, T.T. (2001). Density estimation for spatial linear process. *Bernoulli,* **7**, 657-668.

Marron, J.S. and Ruppert, D. (1994). Transformations to reduce the boundary bias in kernel density estimation. *J. Roy. Statist. Soc.,* **B56**, 653-671.

Newman, C.M. (1980). Normal fluctuations and FKG inequalities. *Comm. Math. Phys.,* **74**, 119-128.

Newman, C.M. (1983). A general central limit theorem for FKG systems. *Comm. Math. Phys.,* **91**, 75-80.

Newman, C.M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables. In: *Inequalities in Statistics and Probability* (Y.L. Tong, ed.), vol. 5, Institute of Mathematical Statistics, California, 1984, 127-140.

Parzen, E. (1962). On estimation of probability density and mode. *Ann. Math. Statist.,* **33**, 1065-1070.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.

Prakasa Rao, B.L.S. and Dewan, I. (2001). Associated sequences and related inference problems. Handbook of Statistics, 19, *Stochastic Processes*: *Theory and Methods* (eds. C.R. Rao and D.N. Shanbhag), North Holland, Amsterdam, 693-728.

Qaim, M. (2003). BT cotton in India: Field trial results and economic projections. *World Development,* **31**, 2115-2127.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of density functions. *Ann. Math. Statist.,* **27**, 832-837.

Roussas, G.G. (1991). Kernel estimates under association: Strong uniform consistency. *Statist. Probab. Lett.,* **12**, 393-403.

Roussas, G.G. (1999). Positive and negative dependence with some statistical applications. In: *Nonparametrics and Time Series*, S. Ghosh, Marcel Dekker (eds.), New York, 757-788.

Scaillet, O. (2004). Density estimation using inverse Gaussian and reciprocal inverse Gaussian kernels. *Jour. Nonpar. Statist.,* **16**, 217-226.

Shaked, M. and Shanthikumar, J. George (2007). *Stochastic Orders.* Springer Series in Statistics, Springer, New York.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, New York.

Wand, M.P., Marron, J.S. and Ruppert, D. (1991). Transformations in density estimation. *Jour. Amer. Statist. Assoc.,* **86**, 343-361.

Yu, H. (1993). A Glivenko-Cantelli lemma and weak convergence for empirical processes of associated sequences. *Prob. Theo. Rel. Fields,* **95**, 357-370.

Zhang, S. (2010). A note on the performance of the gamma kernel estimators at the boundary. *Statist. Probab. Lett.,* **80**, 548-557.